DOCUMENT RESUME

ED 144 952 TM 006 347

AUTHOR Medley, Donald M.
TITLE An Approach to the Definition and Measurement of Teacher Competency.
PUB DATE 7 Apr 77
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; Effective Teaching; Elementary Secondary Education; *Performance Tests; Predictor Variables; Student Teacher Relationship; Teacher Behavior; Teacher Certification; *Teacher Evaluation; *Teacher Qualifications; Teaching Skills; *Test Reliability; *Test Validity

ABSTRACT

Certain classroom behaviors, indicants of teacher effectiveness, were identified by a group of classroom teachers. These behavioral indicators provided the basis for a performance test--in the form of an observation schedule--which can be used for certifying candidates as competent to teach school. Five standardized observation instruments were used in 60 classrooms in an attempt to objectively record teacher behavior and classroom interactions. These included the Coping Analysis Schedule for Educational Settings (CASES), the Spaulding Teacher Activity Rating Schedule (STARS), the Observation Schedule and Record, Form 5, Verbal (OScAR 5V), the Florida Classroom Climate and Control System (FLACCS), and the Teacher Practices Observation Record (TPOR). Items from the five observation instruments which related to the previously identified behavioral indicators were combined to yield an overall score for each area, thus forming the basic design of the new observation schedule. Test reliability and validity data are appended. (MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED144952

AN APPROACH TO THE DEFINITION AND MEASUREMENT OF TEACHER COMPETENCY

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Donald M. Medley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Donald M. Medley
University of Virginia

Presented in a Symposium titled:

Process/Product Research in the Real World: An Attempt to Develop and Validate a Competency Based Teacher Certification System

TM006 347

1977 Annual Meeting
American Educational Research Association
New York, April 7, 1977

AN APPROACH TO THE DEFINITION AND MEASUREMENT OF TEACHER COMPETENCIES

Donald M. Medley
University of Virginia

The specific purpose of the project that is being reported today is to identify observable classroom behaviors whose presence--or absence--in a sample of the performance of a teacher can serve as an indicator of the effectiveness of that teacher. The principal product is to be a performance test that can be used as a basis for certifying candidates as competent to teach school. Physically, the instrument is expected to take the form of an observation schedule that contains a list of indicators of competence (or lack of it). After a reasonable amount of training, an observer will be able to use the instrument to record which of the indicators are present and which of them are absent in a brief sample of classroom behavior. A set of such records of one teacher's behavior can then be scored to yield (1) an estimate of how effective that teacher is likely to be, and (2) a diagnostic profile indicating specific ways in which he might improve.

The test will, therefore, not measure effectiveness directly; rather, it will measure the degree to which a candidate possesses a set of competencies known to characterize effective teachers.

Our search for indicators of competence began with the development by the teachers themselves of a list of competencies they perceived as important to effective teaching, with suggested indicators of the presence of each. After developing measures of each indicator, we observed behavior in a sample of classrooms to estimate the amount of each competence each teacher displayed, measured pupil gains in those classrooms on a variety of measures, and intercorrelated the two in order to find out which competencies did in fact characterize the effective teacher.

My role here today is to describe how we developed measures of each of the competencies on the teachers' list; later members of the panel will describe other phases of the project. I believe that the procedure that I shall describe below is of interest not only as it relates to our project, but as an approach to competency assessment as a part of the evaluation of teacher education programs in general, and competency-based programs in particular. I propose today to describe to you how a set of competencies quite similar to those used to specify the goals of teacher education programs were operationally defined and reliably measured without the use of high-inference ratings, and using only existing and generally available low-inference systems. I shall begin by setting forth a series of operational decisions or working assumptions which together describe the approach that we have used. Then I will describe the analysis of the data we used to show how the measures evolved step by step. Finally I will present and discuss the results of this analysis.

APPROACH

Decision Number One was that the instrument produced would be a low-inference rather than a high-inference one. The main reason for this was a practical one: the greater vulnerability of high-inference ratings to legal attacks in the courts. Sooner or later a candidate who fails to pass the test will decide to go to court and question its validity. If the decision to fail him has been based on high-inference ratings, the defense must be based ultimately on the expertise of the rater, since his judgments in the form of ratings will constitute the sole documentary evidence that the candidate was incompetent. If the decision is based on low-inference records, however, an objective record of what the candidate did and did not do on the examination will be in the record, together with a competency score derived from that record by the use of the same scoring key that was used on every other candidate. The record is there; the behaviors on which the failing score was based can be examined; and it can be shown that any other candidate who had behaved in the same way would also have failed. Thus any charge of bias or of unreliability can be refuted, and a strong case built against any charge of invalidity.

The other advantages of low-inference instruments are well-known and will not be recapitulated here.

Decision Number Two was to try out only indicators that had been identified beforehand as promising, instead of blindly testing all possible indicators. The chances that a behavior which is found to predict teacher effectiveness will stand up on cross

validation should be greater when only a limited number of promising behaviors are tested. In the present instance, records were made of the incidence in each classroom on over 1,300 possible behavioral indicators. Since 58 different measures of outcomes were also obtained, over 80,000 process-product correlations could have been calculated and examined. By carefully selecting fewer than 100 behavioral indicators from the 1,300 to be correlated with outcomes, we reduced the total number of correlations substantially.

Decision Number Three was to have the teachers themselves identify these indicators of competence. Special task forces of classroom teachers from the school system in which the study was conducted spent a year meeting, discussing, and studying various definitions of teacher competence. They finally chose the competencies shown in Exhibit 1. More importantly, they also listed under each area of competency specific teacher and student behaviors whose presence in the classroom should indicate that the teacher possessed each competency.

The competencies listed are not very different from those on other lists that have been developed by other groups as specifications of PBTE programs; what is unusual about this list is that each competency is defined in terms of specific, observable behavioral indicators. It is these indicators that formed the basis for the development of the measures described below, and that will also be the basis of the performance test yet to be developed.

Decision Number Four was not to attempt to construct an observation instrument specially designed to measure these behavioral

indicators. One reason behind this decision was the unanimous opinion of project consultants that building a valid observation system is not as easy as it looks. Too many similar projects have underestimated the time and effort such an enterprise entails; and too many have ended up with a mass of data that are largely uninterpretable. In order to avoid this particular trap, we decided to use existing instruments of proven quality. We finally adopted the five that are listed in Exhibit 2--CASES, FLACCS, OScAR 5V, STARS, and the TPOR. Each of the first four of these instruments is the product of 15 to 20 years of development and refinement. All five of them have been used in more than one previous study by more than one investigator. Taken together, these instruments provided data on 1,348 categories or items of behavior; taken separately, they represented several distinct points of view about classroom behavior and about observational methodology. It seemed to us that almost any indicator we might care to study could be identified on one or more of these instruments (and we found this to be the case).

Decision Number Five was to base each indicator on more than one single item whenever possible. Cronbach (1951) has shown how rapidly a score based on a set of test items with low inter-correlations increases in internal consistency as the number of items increases--a phenomenon familiar to every measurement student as the Spearman-Brown Law.

Although scores based on individual items of behavior are extremely unstable and ambiguous as to meaning, the combination of as few as four or five into a composite can produce a

dramatic increase both in stability and in internal consistency, with an accompanying reduction in ambiguity (Medley and Mitzel, 1963, pp. 303-309).

PROCEDURE

The first step in the process was to carefully scan each instrument for items related to each of the behavioral indicators listed in Exhibit 1. When two or more such items were found on the same instrument they were grouped to form a scale or scoring key to measure that indicator on that instrument.

Not every indicator of competence could be measured on every instrument, but most of them could be scored on at least two. There were a few instances in which a single item corresponded so closely to a competency indicator that it could function as a one-item key.

The second step was to combine all of the single-instrument keys designed to measure the same indicator of competence into a single key that would yield a composite or multi-instrument measure of the behavior in question.

The third and final step was to combine all of the keys scored on competency indicators in each area of competency into a key designed to yield an overall score for each area. In this step, only keys based on teacher behavior were used, because indicators based on student behaviors showed too much overlap between areas.

In the remainder of this presentation I will describe what happened when we implemented this plan. Since the statistics

I shall use in describing what happened are reliability coefficients, I will begin by describing how we estimated the reliability of a competency score.

RELIABILITY ESTIMATION

Records of behavior were made in the classroom of 59 teachers on all five instruments, CASES, FLACCS, OScAR, STARS, and TPOR, with at least two observers recording behavior on each instrument. The number of observers, and the number of visits each observer made to each classroom, varied from one instrument to the next; but every observer using any given instrument visited each classroom and recorded behavior on that instrument at least once, and no two observers visited the same classroom at the same time.

Experience has shown that the scores of a sample of classrooms on the typical item on an observation scale are not normally distributed, but are more likely to fall into a J-shaped curve with a mode near zero. This tends to distort product-moment correlations involving such items. Therefore, the set of mean scores per visit on each item recorded by each observer in the 59 classrooms was normalized by a T transformation. In addition to fitting a normal curve to the distribution for each item, this transformation set the mean and standard deviation of each item for each observer at 50 and ten, approximately.

This had two important additional effects: (1) it sacrificed all information in the item scores except information about

the rank order of the 59 classrooms on that item; and (2) it also eliminated any systematic error in the form of observer bias.

Exhibit 3 shows the analysis of variance design used to estimate the reliability of a key made up of K items scored on records made by R recorders on a single instrument.

In the model for this analysis, Recorders were, of course, regarded as random, as were Classrooms, so that findings could be generalized to other recorders visiting other classrooms.

Since the indicators of each competency suggested by the teachers (See Exhibit 1) were not meant to be exhaustive, but only samples which could have been replaced with other specific indicators, Items were also regarded as random, so that errors of measurement due to inconsistencies between scores based on different items are included in the measurement errors allowed for by the reliability coefficients as defined in Exhibit 3.

The reliability coefficient of a measure as scored may be interpreted as the correlation between a set of scores based on the total (or mean) of the S records per classroom made on K items on the key by the R recorders acually used, and a set of scores based on the total (or mean) of S other records per classroom on K other items by R other recorders visiting the same classrooms at different times. This is a conservative estimate of reliability, since errors due to heterogeneity of items, instability of classroom behavior, and observer errors all enter into the error of measurement and can lower the reliability

coefficient. Very few of the reliability coefficients of observational measurements reported in the literature take account of item heterogeneity; and by no means all of them allow for instability of behavior. Too many are mere coefficients of observer agreement.

The coefficient of reliability just defined is a good index of the accuracy of scores on a measure, and will be used for this purpose. It is not, however, very useful for comparing keys made up of different items. For this purpose we will use a statistic we will call the reliability per iter. This coefficient estimates the correlation between a set of scores recorded on a different item by a different recorder visiting the same classrooms but at other times. It is assumed that both items are randomly chosen.

Since it is independent of the number of items and recorders, the reliability per item may be taken as an estimate of the degree to which the items on the key are saturated with a single common factor. The higher this coefficient, the less ambiguity in a score. In particular, if a key is lengthened by the addition of items equivalent to those already on the key, the reliability of the scale increases. If non-equivalent items are added, the reliability may or may not increase, but the reliability per item should decrease.

Exhibit 4 shows the design used to estimate the reliability of a composite measure--that is, one scored across two or more different instruments, In this model, Instruments is regarded as a fixed effect, since there is no intention to generalize to

other instruments. The only added complexity that arises in estimating reliabilities of composites is a consequence of the fact that the number of items and the number of recorders varies from one instrument to another. This fact was taken into account so that the measures of coefficients from both designs are the same.

RESULTS

Before commenting on the findings as a whole, let me discuss an example in detail.

On the first page of Exhibit 5, the first indicator listed under Area 3: Effective Communication, is labeled 3A-T. This indicator has two single-instrument measures. One comes from OScAR and has 10 items; the other consists of a single item on the TPOR. The indicator being measured. incidentally, is whether the teacher gives clear and explicit instructions which are understood by the students.

TPOR happened to contain one item which required the observer to record that particular behavior, so this item was scored as a one-item scale. Scores on the item all have a reliability of .20, and a reliability per item of .07 according to the Exhibit. It may seem odd that the reliability per item is smaller than the reliability of the one item; this reflects the fact that the reliability per item is based on one record only while the reliability is based on more than one.

OScAR contains ten items in the form of two-event sequences which seem to be instances of clear explaining (or its opposite).

One example which may be identified in an OScAR record is an instance in which a pupil asks a question immediately after the teacher has given information. Such an occurence is regarded as an indication of absence of the competency. Frequencies on ten such items form the basis for the OScAR measure of indicator 3 A-T. The reliability of such scores was estimated to be .69; the reliability per item was estimated to be .11.

When the TPOR item and the 10 OScAR items were combined into an 11-item composite, the reliability of the composite was estimated to be .67 and the reliability per item was estimated to be .09. In this instance, both coefficients were lower in the composites. This is not typical.

If you will turn to the first page of Exhibit 7, you can see what happened when 3A-T was combined with 3B-T and 3C-T to get an area score--a measure of how competent a teacher is in communicating effectively with students. The reliability of this measure is .95, and its reliability per item is .08.

Exhibit 5 lists results obtained on 49 single-instrument keys (some of which appear more than once) based on teacher behaviors, and 14 composites across two or more instruments. Exhibit 6 lists results for 45 single-instrument keys based on student behaviors, and 13 composites. And Exhibit 7 lists results for eight competency area keys based on teacher behaviors only.

The median reliability of the single-instrument keys based on teacher behavior was .70, and the median reliability

of the single-instrument keys based on student behavior was .51.

The median reliability of the multi-instrument composite keys was .88 for teacher behaviors and .67 for student behaviors.

The median reliability of the 3 competency keys was .94. Seven of the eight had reliabilities above .80.

The median reliabilities per item are as follows: for single-instrument keys, .07; for composites, .08; and for competency areas, .08.

DISCUSSION

In the past I have always thought of reliability as something you worry about when you have no way of establishing the validity of a test. And the validity evidence that Drs. Soar and Lorentz present below constitute the really important results of this study. Despite this, the reliability data in Exhibits 5, 6, and 7 have importance independent of their validity as predictors of outcomes. They have important implications for us in our efforts to construct a performance test for teachers and also for teacher educators interested in program evaluation.

Before one can attempt to do either of these things--build a test or evaluate a teacher education program--he must commit himself to a set of performance competencies, of teacher behaviors, that his test is supposed to measure or his program is supposed to develop. A teacher performance test must measure some kind of behavior emitted by the teacher who takes the test. A teacher education program must teach teachers some specific ways

of behaving in the classroom. In each instance, the behaviors--or competencies, as they are called nowadays--are chosen because possession of them is expected to enhance a teacher's effectiveness. Whether it does or not is an empirical question, and the process by which that question is answered is the process of validation. Validation usually consists in showing that scores on the competencies correlate with pupil gains in the teachers' classes.

But there is another question, scarcely less important than the validity question; and that is whether the test really measures the competencies, or whether the program actually develops them in its graduates. Reliability evidence such as I have presented is relevant to this question and therefore has an importance of its own quite apart from what I have called the validity question.

When a competence is defined in terms of specific indicators visible in the behavior of a competent teacher, as was the case in this project; and when observers record the extent to which these indicators appear in the behavior of a set of teachers, we have information from which answers can be derived. If independent observers agree on how often the behaviors occur in each classroom; if stable differences are observed between frequencies of the behaviors in the classrooms of different teachers; and if the different behavioral indicators are consistent with each other; then it seems safe to infer that the specified competency has been measured.

If a competency is reliably measured and its correlation

with pupil gain measures is low or negative, there is reason to omit that competency from the test specifications or list of program objectives. If a competency measure is not reliably measured, no meaningful validity data can be obtained. It is in this respect that reliability data about competency measures are intrinsically important.

The substantial reliabilities found for single-instrument measures--a median of .70 for measures based on teacher behaviors--are impressive mainly as indicating that existing instruments can be used to measure indicators of competency, even though they were not constructed for this purpose. Our expectation that well-established instruments, ones that had been developed and refined over a period of years, would contain reliable items relevant to many areas of competence was supported by these results.

The fact that when items from quite different instruments were combined into composite measures the composites were more reliable than the single-instrument measures would seem to provide additional support for the conclusion that the specified behaviors are in fact being measured. The median reliability for teacher-based composites of .88 is high enough for most purposes to which the proposed test would be used, and more than adequate for program evaluation.

The extremely high reliabilities obtained when all indicators of a single competency were combined into one key strain credulity. We have checked these figures often enough to be sure that they are correct; but they are hard to believe.

So far as I know, this is the first time anyone has developed scoring keys based on such large numbers of low-inference items. Knowing that the average item intercorrelations are of about the same order of magnitude as those between objective test items should lead one to expect comparable reliabilities on measures of comparable length; but we did not expect them.

The approach used in developing these measures intentionally avoided any use of empirical data as we combined items into longer and longer keys, so we have not looked at intercorrelations. Perhaps it is time we did. I strongly suspect that there is a common facter running through all of these measures, analogous to the "g" factor than loads all cognitive items. Could it be teacher competence?

By way of summing up, let me observe that if there is one thing we have found out it is that if you want to measure some set of teacher competencies you are interested in, there seems to be no great problem. Just use several low-inference observational systems at least as good as those we used, and you can do it.

I have already pointed out the implications for teacher evaluation and program evaluation; let me mention one implication for research. It is time someone collected a set of data like ours, and then scored the records on all of the competencies that previous research indicates may be correlated with outcomes, and so replicates all of the research in a single study.

EXHIBIT I

COMPETENCY AREAS IDENTIFIED BY TEACHER TASK FORCE
WITH BEHAVIORAL INDICATORS

Teacher Competency Area	Teacher Behaviors (Process)	Student Outcomes
1) Gathers and uses information relating to individual differences among students	a) Maintains and uses formal/ informal up-to-date records on individual students	a) working at task on appropriate level of instruction.
	b) Consults appropriate authorities to select and administer appropriate standardized tests when information is needed on individual students and follows through with information on learning levels, interests, values, cultural and socio-economic background.	b) reduced deviant behavior
	c) Teacher recognizes limitations and seeks additional professional help.	c) better physical,mental health
2) Organizes pupils, resources and materials for effective instruction	a) Selects goals & objectives appropriate to pupil need	a) enjoys class, happy, smiles
	b) Matches student with appropriate material	b)c) on task, actively involved
	c) Gathers multi-level materials	b)c) evidences academic growth
	d) Teacher involves student in organizing and planning	d) absence of withdrawn behavior
		e) enthusiastically involved
		f) evidence of involvement
3) Demonstrates ability to communicate effectively with students	a) Gives clear explicit directions which are understood by students	a) less confusion, less time wasted
	b) Pauses, elicits and responds to student questions before proceeding.	b) more relaxed, less frustration
	c) Uses a variety of methods, verbal and non-verbal, to deliver instructions	c) self-directed to move toward task

EXHIBIT I (Continued)

Teacher Competency Area	Teacher Behaviors (Process)	Student Outcomes
4) Assists students in using a variety of relevant communication techniques.	a) Demonstrates proper listening skills	a) acquires capacity to be a good listener
	b) Respects individual's right to speak	b) students able to speak freely
	c) Utilizes non-verbal communication skills	c) able to follow directions, on task
	d) Utilizes written language as type of communication	d) discriminates acceptable or not acceptable behavior
		e) able to communicate through writing
5) Assists students in dealing with their misconceptions or confusions, using relevant clues and techniques	a) Utilizes student feedback, verbal and non-verbal, to modify own teaching behavior	a) Students ask questions
	b) Demonstrates flexibility in classroom management practices	b) students feel free to interrupt presentations
	c) Provides opportunity for student-initiated questions	c) movement toward task
	d) When student not on task, teacher makes contact	
6) Responds appropriately to coping behavior of students	a) Maintains self-control in various classroom situations and interactions with students	a) Absence of student manipulation
	b) Recognizes and treats individual student behavior	b) modifies behavior positively
	c) Seeks appropriate help from others	c) reduction of disruptive behavior
	d) Accepts necessity of dealing with individual students on an individual basis	
7) Uses a variety of methods and materials to stimulate & promote pupil learning	a) Uses more than one teaching method in a single presentation	a) attending (attentive) behavior
	b) Adapts methods & materials to instructional situation and to established goals & objectives	b) motivated
		c) actively involved

EXHIBIT I (Continued)

Teacher Competency Area	Teacher Behaviors (Process)	Student Outcomes
8) Promotes self-awareness and positive self-concepts in students	a) Provides opportunity for each student to meet success daily.	a) student working on individual level
	b) Provides variety of materials	b) moving toward self-direction, attending to task
	c) Evidence of a personal one-to-one relationship with each student	c) knowledge of variety of cultural & socio-economic background
	d) Provides opportunity for student to have voice in decision making	d) evidence of importance as class member--group involvement
	e) evidence of praise and/or rewards in operation	e) assumes responsibility for own success or failure
	f) Supportive classroom management	f) evidence of enthusiasm
9) Reacts with sensitivity to the needs and feelings of others	a) Accepts & incorporates student ideas	a) expresses ideas & opinions different to those of teacher or peers
	b) Listens to students and provides feedback	b) high interest
	c) Evidence of an opportunity for one-to-one counseling and absence of evidence that students are rejected (brushed off)	c) student/teacher rapport is evident, develops sense of belonging, evidence of confidence in teacher
10) Engages in personal and professional growth	a) Reads widely and critically	
	b) Maintains membership & active participation in professional organizations appropriate to the individual	
	c) Exchanges ideas & teaching techniques with colleagues	
	d) Continuously improves knowledge and skills	
	e) Engages in continuous self-evaluation	

EXHIBIT I (Continued)

Teacher Competency Area	Teacher Behaviors (Process)	Student Outcomes
11) Works effectively with pupils, parents, colleagues, community & educational administrators of school system	a) Attends school-related community activities b) Accepts responsibility for some community activity c) Supportive of school policies.	

EXHIBIT 2

OBSERVATION INSTRUMENTS

The Competency Based Teacher Certification Project at West Georgia College is using five observation instruments in 60 on-going Carroll County classrooms in an attempt to objectively record teacher behavior and classroom interactions.

The Coping Analysis Schedule for Educational Settings (CASES) was designed to measure pupil socialization and consists of 13 categories of "coping" behaviors which are identified by descriptive statements.

The Spaulding Teacher Activity Rating Schedule (STARS) is a category system which examines the cognitive instructional strategies of teachers as well as their affective and control techniques. STARS consists of 25 categories which are subsumed under the subtitles of affective behavior, motor and social structuring, concept attainment, concept checking, and value expression.

The Observation Schedule and Record, Form 5, Verbal (OScAR 5V) concentrates on the verbal behavior of teachers. It consists of 14 categories designed to measure questioning and feedback strategies, in addition to four categories for pupil-initiated utterances.

The Florida Classroom Climate and Control System (FLACCS) examines the control tactics of teachers as well as their affective behaviors. It includes items relating to the nature of classroom structure, teacher and student control strategies, and teacher and student affective behaviors, both positive and negative.

The Teacher Practices Observation Record (TPOR) measures the congruency of observed teacher behavior in the classroom with educational practices advocated by John Dewey, and consists of 62 items which describe teacher behavior. In relation to Dewey's philosophy of experimentalism, 31 of the items are positive and 31 are negative. All even-numbered items are pro-Dewey and all odd-numbered items are non-Dewey.

It is hoped that by utilizing several instruments, with their inherent emphases and approaches to classroom behavior, we will have a broad data base from which to draw inferences.

EXHIBIT 3

DESIGN FOR RELIABILITY ANALYSIS OF A SCALE SCORED AS ONE INSTRUMENT

Source of Variation	Degrees of Freedom	Mean Squares	
		Observed	Expected
Between classes	58	a	$KR\sigma^2_c + R\sigma^2_{ck} + \sigma^2$
Interaction, class x item	58(K-1)	c	$R\sigma^2_{ck} + \sigma^2$
Residual	58K(R-1)	d	σ^2
Total Variation	58S		
Fitting S means*	S		
Total	59S		

59 = Number of classrooms

K = Number of items

R = Number of recorders

S = KR

Estimation: σ^2 (=) d

σ^2_{ck} (=) $\frac{c - d}{R}$

σ^2_c (=) $\frac{a - c}{S}$ (c > d)**

Reliability: $$\frac{\sigma^2_c}{\sigma^2_c + \frac{\sigma^2_{ck}}{K} + \frac{\sigma^2}{S}}$$

Reliability per item: $$\frac{\sigma^2_c}{\sigma^2_c + \sigma^2_{ck} + \sigma^2}$$

*No main effect for items is shown because the normalization of each item makes all item means the same.

**When $c \leq d$, c and d are pooled to yield an estimate e of σ^2 (with 58(S-1) degrees of freedom), and e is substituted for c.

EXHIBIT 4

DESIGN FOR RELIABILITY ANALYSIS OF A SCALE SCORED ACROSS DIFFERENT INSTRUMENTS

Source of Variation	Degrees of Freedom	Mean Squares	
		Observed	Expected
Between classes	58	a	$S\sigma^2_c + \bar{R}\sigma^2_{ck} + \sigma^2$
Interaction, class x instrument	58(I-1)	b	$\bar{P}\sigma^2_{ci} + \bar{Q}\sigma^2_{ck} + \sigma^2$
Interaction, class x item (in instrument)	58(K-I)	c	$\bar{R}\sigma^2_{ck} + \sigma^2$
Residual	58(S-K)	d	σ^2
Total Variation	58S		
Fitting S means	S		
Total	59S		

59 = Number of classrooms

I = Number of instruments

K_I = Number of items on instrument I $\quad (K = \Sigma K_I)$

R_I = Number of recorders using instrument I $\quad (R = \Sigma R_I)$

$S_I = K_I R_I$ = Number of records made on instrument I $\quad (S = \Sigma S_I)$

$$\bar{P} = \frac{1}{I - 1}\left(S - \frac{\Sigma S_I^2}{S}\right)$$

$$\bar{Q} = \frac{1}{I - 1}\left(R - \frac{\Sigma R_I S_I}{S}\right)$$

$$\bar{R} = \frac{1}{K - I}(S - R) \qquad \bar{K} = \frac{S}{\bar{R}}$$

Estimation: $\sigma^2 \ (=) \ d$

$\sigma^2_{ck} \ (=) \ \frac{c - d}{\bar{R}}$

$\sigma^2_c \ (=) \ \frac{a - c}{S}$ (Items Random)

$$\text{Reliability} = \frac{\sigma^2_c}{\sigma^2_c + \frac{\sigma^2_c}{\bar{K}} + \frac{\sigma^2}{S}}$$

$$\text{Reliability per item:} \quad \frac{\sigma^2_c}{\sigma^2_c + \sigma^2_{ck} + \sigma^2}$$

EXHIBIT 5

ESTIMATES OF RELIABILITIES OF COMPETENCY INDICATORS
BASED ON TEACHER BEHAVIORS

(Number of Classrooms = 59)

Competency Indicator	Instrument	Number of Items	Reliability per Item	Reliability Total
AREA 2:				
2A-T	TPOR	3	.00	.00
2B-T	TPOR	3	.00	.00
2C-T	TPOR	1	.15	.35
2D-T	OScAR	7	.01	.11
AREA 3:				
3A-T	OScAR	10	.11	.69
	TPOR	1	.07	.20
	Composite	11	.09	.67
3B-T	OScAR	6	.05	.39
	OScAR	4	.02	.16
	FLACCS	2	.13	.38
	FLACCS	2	na*	na*
	STARS	40	.07	.87
	Composite	54	.07	.91
3C-T	FLACCS	20	.08	.79
	STARS	29	.10	.88
	Composite	49	.09	.92
AREA 4:				
4A-T	OScAR	20	.10	.81
	FLACCS	1	.09	.16
	STARS	30	.07	.84
	Composite	50**	.08	.91
4B-T	OScAR	20	.10	.81
	TPOR	5	.06	.47
	STARS	52	.08	.91
	Composite	76**	.08	.94
4C-T	FLACCS	8	.05	.43
AREA 5:				
5A-T	STARS	42	.07	.87
5B-T	OScAR	3	.00	.02
	STARS	27	.07	.85
	Composite	30	.07	.86
5D-T	STARS	43	.09	.92
5E-T	FLACCS	6	.10	.56
5F-T	TPOR	5	.03	.34

EXHIBIT 5 (continued)

Competency Indicator	Instrument	Number of Items	Reliability per Item	Reliability Total
AREA 6:				
6A-T	OScAR	5	.02	.35
	FLACCS	9	.19	.80
	TPOR	1	.01	.03
	STARS	27	.07	.85
	Composite	44**	.08	.90
6B-T	FLACCS	2	.13	.38
	STARS	27	.07	.85
	Composite	29	.07	.86
6D-T	TPOR	6	.00	.00
	STARS	27	.07	.85
	Composite	33	.07	.86
AREA 7:				
7A-T	OScAR	41	.05	.82
7C-T	FLACCS	1	.61	.76
AREA 8:				
8C-T	FLACCS	7	.18	.76
	STARS	31	.08	.87
	Composite	39**	.08	.90
8E-T	FLACCS	2	na*	na*
	TPOR	1	.21	.45
	Composite	3	.03	.18
8F-T	FLACCS	15	.11	.80
	FLACCS	5	na*	na*
	TPOR	2	.00	.00
	Composite	22	.08	.80
AREA 9:				
9A-T	TPOR	3	.07	.42
9B-T	OScAR	35	.07	.86
	FLACCS	5	.20	.71
	STARS	41	.07	.87
	Composite	81	.07	.93
9C-T	FLACCS	3	na*	na*
	TPOR	1	.00	.00
	STARS	21	.05	.72
	Composite	25	.06	.77

* Reliability data not available for this key.

** The discrepancy in item numbers is due to minor revisions in scoring keys between steps.

EXHIBIT 6

ESTIMATES OF RELIABILITIES OF COMPETENCY INDICATORS
BASED ON STUDENT BEHAVIORS

(Number of Classrooms = 59)

Competency Indicator	Instrument	Number of Items	Reliability per Item	Reliability Total
AREA 1:				
1B-S	FLACCS	7	.07	.52
	TPOR	1	.00	.00
	Composite	8	.08	.60
1C-S	FLACCS	6	.03	.30
AREA 2:				
2A-S	FLACCS	2	.00	.00
2B/C-S	FLACCS	6	.04	.29
	TPOR	2	.26	.67
	CASES	6	.11	.72
	Composite	15**	.10	.80
2D-S	FLACCS	4	.00	.00
	CASES	12	.08	.80
	Composite	17**	.06	.79
2E-S	CASES	5	.11	.67
2F-S	CASES	6	.01	.13
AREA 3:				
3A-S	CASES	7	.02	.32
3C-S	FLACCS	4	.13	.54
	CASES	5	.08	.61
	Composite	10**	.09	.74
AREA 4:				
4B-S	OScAR	20	.10	.81
	OScAR	18	.04	.64
	FLACCS	7	.02	.19
	CASES	3	.11	.54
	Composite	49**	.07	.89
4C-S	FLACCS	3	.00	.00
	CASES	6	.01	.13
	Composite	9	.01	.24
4D-S	FLACCS	8	.07	.57
AREA 5:				
5A-S	OScAR	20	.04	.61
	CASES	2	.00	.00
	Composite	22	.05	.68

EXHIBIT 6 (continued)

Competency Indicator	Instrument	Number of Items	Reliability per Item	Reliability Total
5B-S	FLACCS	1	.23	.38
5C-S	FLACCS	5	.20	.71
	CASES	6	.01	.13
	Composite	12**	.05	.60
AREA 6:				
6A-S	CASES	2	.07	.47
6B-S	CASES	7	.02	.32
6C-S	OScAR	2	.30	.57
	OScAR	1	.35	.68
	CASES	8	.08	.74
	Composite	11	.10	.82
AREA 7:				
7A-S	CASES	2	.02	.10
7C-S	TPOR	2	.26	.67
	CASES	4	.07	.48
	Composite	6	.09	.65
AREA 8:				
8B-S	TPOR	1	na*	na*
	CASES	2	.00	.00
	Composite	3	.00	.00
8C-S	CASES	6	.01	.13
8E-S	FLACCS	1	.07	.12
	CASES	3	.12	.54
	Composite	4	.09	.53
8G-S	FLACCS	2	.39	.72
	CASES	4	.07	.48
	Composite	6	.09	.64
AREA 9:				
9A-S	TPOR	3	.00	.00
9B-S	FLACCS	1	.54	.70
	CASES	4	.07	.48
	Composite	5	.09	.57
9D-S	OScAR	12	.05	.63

EXHIBIT 7

ESTIMATES OF RELIABILITIES OF AREAS OF TEACHER COMPETENCY

(Number of Classrooms = 59)

Competency Indicator	Instrument	Number of Items	Reliability per Item	Reliability Total
AREA 2: Organizing for Instruction				
2A-T	TPOR	3	.00	.00
2B-T	TPOR	3	.00	.00
2C-T	TPOR	1	.15	.35
2D-T	OScAR	7	.01	.11
Area Score		14	.03	.50
AREA 3: Effective Communication				
3A-T	OScAR	10	.11	.69
	TPOR	1	.07	.20
3B-T	OScAR	6	.05	.39
	OScAR	4	.02	.16
	FLACCS	2	.13	.38
	FLACCS	2	na*	na*
	STARS	40	.07	.87
3C-T	FLACCS	20	.08	.79
	STARS	29	.10	.88
Area Score		98**	.08	.95
AREA 4: Using Variety of Communication Techniques				
4A-T	OScAR	20	.10	.81
	FLACCS	1	.09	.16
	STARS	30	.07	.84
4B-T	OScAR	20	.10	.81
	TPOR	5	.06	.47
	STARS	52	.08	.91
4C-T	FLACCS	8	.05	.43
Area Score		85**	.08	.94
AREA 5: Using Relevant Clues and Techniques to Clarify Student Misconcepti				
5A-T	STARS	42	.07	.87
5B-T	OScAR	3	.00	.02
	STARS	27	.07	.85
5D-T	STARS	43	.09	.92
5E-T	FLACCS	6	.10	.56
5F-T	TPOR	5	.03	.34
Area Score		86**	.08	.95

EXHIBIT 7 (continued)

Competency Indicator	Instrument	Number of Items	Reliability per Item	Reliability Total
AREA 6: Appropriate Response to Coping Behavior of Students				
6A-T	OScAR	5	.02	.35
	FLACCS	9	.19	.80
	TPOR	1	.01	.03
6B-T,6D-T-	STARS	27	.07	.85
6B-T	FLACCS	2	.13	.38
6D-T	TPOR	6	.00	.00
Area Score		52**	.08	.91
AREA 7: Dealing with Methods and Materials for Pupil Learning				
7A-T	OScAR	41	.05	.82
7C-T	FLACCS	1	.61	.76
Area Score		42	.06	.83
AREA 8: Dealing with Students' Self-Awareness and Self-Concept				
8C-T	FLACCS	7	.18	.76
	STARS	31	.08	.87
8E-T	FLACCS	2	na*	na*
	TPOR	1	.21	.45
8F-T	FLACCS	15	.11	.80
	FLACCS	5	na*	na*
	TPOR	2	.00	.00
Area Score		61**	.08	.93
AREA 9: Dealing with Sensitivity to Needs and Feelings of Others				
9A-T	TPOR	3	.07	.42
9B-T	OScAR	35	.07	.86
	FLACCS	5	.20	.71
	STARS	41	.07	.87
9C-T	FLACCS	3	na*	na*
	TPOR	1	.00	.00
	STARS	21	.05	.72
Area Score		87**	.07	.94

* Reliability data not available for this key.

** The discrepancy in item numbers is due to minor revisions in scoring keys between steps.